Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2016

Methods for handling missing variables in risk prediction models

Held, Ulrike ; Kessels, Alfons ; Garcia Aymerich, Judith ; Basagaña, Xavier ; Ter Riet, Gerben ; Moons, Karel G M ; Puhan, Milo Alan

Abstract: Prediction models should be externally validated before being used in clinical practice. Many published prediction models have never been validated. Uncollected predictor variables in otherwise suitable validation cohorts are the main factor precluding external validation. We used individual patient data from 9 different cohort studies conducted in the United States, Europe, and Latin America that included 7,892 patients with chronic obstructive pulmonary disease who enrolled between 1981 and 2006. Data on 3-year mortality and the predictors of age, dyspnea, and airflow obstruction were available. We simulated missing data by omitting the predictor dyspnea cohort-wide, and we present 6 methods for handling the missing variable. We assessed model performance with regard to discriminative ability and calibration and by using 2 vignette scenarios. We showed that the use of any imputation method outperforms the omission of the cohort from the validation, which is a commonly used approach. Compared with using the full data set without the missing variable (benchmark), multiple imputation with fixed or random intercepts for cohorts was the best approach to impute the systematically missing predictor. Findings of this study may facilitate the use of cohort studies that do not include all predictors and pave the way for more widespread external validation of prediction models even if 1 or more predictors of the model are systematically missing.

DOI: https://doi.org/10.1093/aje/kwv346

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-126161
Journal Article
Published Version

Originally published at:
Held, Ulrike; Kessels, Alfons; Garcia Aymerich, Judith; Basagaña, Xavier; Ter Riet, Gerben; Moons, Karel G M; Puhan, Milo Alan (2016). Methods for handling missing variables in risk prediction models. American Journal of Epidemiology, 184(7):545-551.
DOI: https://doi.org/10.1093/aje/kwv346

American Journal of Epidemiology Advance Access published September 14, 2016

American Journal of Epidemiology
© The Author 2016. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
DOI: 10.1093/aje/kwv346

Practice of Epidemiology

Methods for Handling Missing Variables in Risk Prediction Models

Ulrike Held*, Alfons Kessels, Judith Garcia Aymerich, Xavier Basagaña, Gerben ter Riet, Karel G. M. Moons, and Milo A. Puhan, for the International COPD Cohorts Collaboration Working Group

* Correspondence to Dr. Ulrike Held, Horten Centre for Patient-Oriented Research and Knowledge Transfer, University of Zurich, Pestalozzistrasse 24, 8032 Zurich, Switzerland (e-mail: Ulrike.Held@usz.ch).

Initially submitted March 9, 2015; accepted for publication December 9, 2015.

Prediction models should be externally validated before being used in clinical practice. Many published prediction models have never been validated. Uncollected predictor variables in otherwise suitable validation cohorts are the main factor precluding external validation. We used individual patient data from 9 different cohort studies conducted in the United States, Europe, and Latin America that included 7,892 patients with chronic obstructive pulmonary disease who enrolled between 1981 and 2006. Data on 3-year mortality and the predictors of age, dyspnea, and airflow obstruction were available. We simulated missing data by omitting the predictor dyspnea cohort-wide, and we present 6 methods for handling the missing variable. We assessed model performance with regard to discriminative ability and calibration and by using 2 vignette scenarios. We showed that the use of any imputation method outperforms the omission of the cohort from the validation, which is a commonly used approach. Compared with using the full data set without the missing variable (benchmark), multiple imputation with fixed or random intercepts for cohorts was the best approach to impute the systematically missing predictor. Findings of this study may facilitate the use of cohort studies that do not include all predictors and pave the way for more widespread external validation of prediction models even if 1 or more predictors of the model are systematically missing.

COPD; decision support techniques; logistic models; meta-analysis; missing data; validation studies

Abbreviations: AUC, area under the receiver operating characteristic curve; COPD, chronic obstructive pulmonary disease; FEV_1, forced expiratory volume in 1 second; IPD, individual patient data; MRC, Medical Research Council; PLATINO, Proyecto Latinoamericano de Investigación en Obstrucción Pulmonar; SEPOC, Quality of Life of Chronic Obstructive Pulmonary Disease Study.

A common barrier to the use or validation of existing prediction models in populations that are different from the populations in which the model was developed is the unavailability of 1 or more predictors in the external data set. For example, the regression model underlying the BODE index (so named because it incorporates body mass index, airflow obstruction, dyspnea, and exercise capacity) to predict mortality in patients with chronic obstructive pulmonary disease (COPD) was not externally validated until data on all predictors (including 6-minute walk distance), as well as outcome, were available from cohort studies from Spain (Phenotype and Course of COPD) and Switzerland (Barmelweid cohort) (1).

Missing data are a serious problem in the verification of prediction models and are a known source of biased results (2). The problem exists when data are missing for individual patients or when predictor data have not been collected in a cohort at all. A simple, but not recommended, approach to handling missing data is to exclude the incomplete patient record, or (in the case of a systematically missing predictor) the entire cohort, from the analysis. Doing so does not address the mechanism that generated the missing data. In a recent review of methodological conduct of validation studies of prediction models, Collins et al. (3) found that inappropriate handling and acknowledgment of missing predictor data from individuals is a common problem in validation studies. A substantial body of literature about imputation techniques for analyses using single and multiple studies emerged recently (4–7). If predictors have not been collected at all—if

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

data for a specific predictor are missing entirely—the problem is even greater. Ahmed et al. (8) demonstrated that uncollected predictors often lead to the omission of entire studies within individual patient data (IPD) meta-analysis. Resche-Rigon et al. (9) addressed the problem of cohort-wide missing variables with a random-effects Cox model for multiple imputation. The use of random-effects models in multiple imputation is rather uncommon, as a recent review by Ahmed et al. (8) showed; they found that in 10 of 15 IPD meta-analyses, the authors ignored the clustering of patients within studies, and in 3 studies, the authors accounted for clustering by including fixed study effects. Jolani et al. (10) addressed the problem of systematically missing predictors by including random effects for predictor variables to allow for between-study heterogeneity. The aim of this paper is to explore how systematically missing predictors could be imputed, and how the model's predictive performance changes depending on different imputation methods as well as cohort-specific characteristics.

Based on an approach by Puhan et al. (11), IPD information on age, dyspnea, and forced expiratory volume in 1 second (FEV_1) (the components of the ADO index) of COPD patients, as well as 3-year mortality rate and sex, was gathered from 9 cohorts, including the Barmelweid cohort (Switzerland) (1), Basque study (Spain) (12) , Copenhagen City Heart Study (Denmark) (13), Cardiovascular Health Study (United States) (14), Jackson Heart Study (United States) (15), National Emphysema Treatment Trial (United States) (16), Phenotype and Course of COPD (Spain) (17), Proyecto Latinoamericano de Investigación en Obstrucción Pulmonar (PLATINO) (South America) (18), and Quality of Life of Chronic Obstructive Pulmonary Disease Study (SEPOC) (Spain) (19). We derived and compared several methods for handling a systematically missing predictor using the full data set as the benchmark. We hypothesized that if variables of the prediction model were missing entirely in some cohorts, some form of imputation would preserve the predictive accuracy of the model better than elimination of the entire cohort would.

METHODS

The original data pool for this analysis consisted of 9 cohort studies that included 7,892 patients without any missing information for the variables of age, modified Medical Research Council (MRC) dyspnea scale grade, percent predicted FEV_1, and survival status within 3 years. We selected the modified MRC dyspnea scale as the predictor to be excluded. The modified MRC dyspnea scale has 3 grades: 0 or 1, 2 or 3, and 4. Age was measured continuously in years, and FEV_1 was dichotomized into less than or equal to 50% and more than 50%. Patients with a FEV_1 of 50% or lower were considered to have severe COPD. Our purpose was not to provide yet another validation of the ADO index, about which we have published previously (11). Instead, we aimed to address the methodological question of how to impute entire predictors if they are missing from a cohort's data. Therefore, we took the liberty of modifying the data set in order to make it fit for the purpose of this analysis. Categories of MRC dyspnea scale grade and FEV_1 were introduced because doing so enhanced the chance of finding differences between imputation methods, thereby leaving more room for improvement compared with the continuous versions of the predictors. The cohorts were divided into 2 categories according to setting: specialized pulmonary medicine setting (Barmelweid, National Emphysema Treatment Trial, and SEPOC) and primary care/nonspecialized care setting (Basque study, Copenhagen City Heart Study, Cardiovascular Health Study, Jackson Heart Study, Phenotype and Course of COPD, and PLATINO), as described elsewhere (11).

Our benchmark model was a logistic regression model with random intercepts by study. Specifically, our model was formulated as $\eta = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, where η is the logit of 3-year mortality; $\mathbf{X}$ is the matrix of predictors that includes a vector of ones (intercept) and the variables age, setting, MRC dyspnea scale grade, and FEV_1; $\boldsymbol{\beta}$ is a vector of regression coefficients associated with the predictors; $\mathbf{Z}$ is a matrix of dummy variables indicating the different studies; $\mathbf{u}$ is a vector of study-specific random effects, with $E(\mathbf{u}) = 0$ and $\mathrm{Var}(\mathbf{u}) = \sigma_u^2$; and $\boldsymbol{\varepsilon}$ is a vector of residuals with $E(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2$.

When using multiple imputation to deal with the missing data, we set the number of imputations as $m = 5$ (20) because the maximum percentage of missing values was below 30%. To simulate a systematically missing predictor, we left out MRC dyspnea scale grade cohort-wide.

We then chose 6 different modeling approaches to evaluate which approach dealt best with the completely missing predictor. These approaches included reestimation of the model parameters to measure the influence of computation techniques rather than model refitting alone. The modeling approaches are listed in detail below.

1a. Exclude the study and use the study-specific setting: Fit the model to include information on specialized/nonspecialized care setting on all available data ($n < 7{,}892$).

1b. Exclude the study and use random study effects: Fit the model with a random study effect on all available data ($n < 7{,}892$).

2. Exclude the predictor: Fit the model without the missing predictor on all data ($n = 7{,}892$).

3. Use mode imputation: Before fitting the model, replace the MRC dyspnea scale grade in the excluded cohort with the mode ($n = 7{,}892$).

4. Fit a multiple-imputation model for MRC dyspnea scale grade with fixed study effects: Conduct multiple imputation based on the variables death, age, FEV_1, sex, and study. Model coefficients for prediction included age, MRC dyspnea scale grade, FEV_1, and fixed study effects ($n = 7{,}892$).

5. Fit a multiple-imputation model for MRC dyspnea scale grade with random study effects: Conduct multiple imputation as described above. Model coefficients for prediction included age, MRC dyspnea scale grade, FEV_1, and random study effects ($n = 7{,}892$).

We repeated this process 9 times to simulate the situation with a missing predictor variable for each of the cohorts. The approaches were compared with respect to different parameters: We measured the predictive accuracy of the models using the area under the receiver operating characteristic curve (AUC) to evaluate discriminative ability. The Wilcoxon test was

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

Table 1. Summary of Number of Patients and 3-Year Mortality Rate in 9 Chronic Obstructive Pulmonary Disease Cohort Studies Conducted in the United States, Europe, and Latin America, 1981–2006

Cohort	No. of Patients	No. of Deaths	3-Year Mortality, %
BMW	231	79	34
Basque study	106	16	15
CCHS	2,273	184	8
CHS	2,109	169	8
JHS	419	29	7
NETT	1,934	516	27
PAC-COPD	330	39	12
PLATINO	173	16	9
SEPOC	317	61	19

Abbreviations: BMW, Barmelweid cohort; CCHS, Copenhagen City Heart Study; CHS, Cardiovascular Health Study; COPD, chronic obstructive pulmonary disease; JHS, Jackson Heart Study; NETT, National Emphysema Treatment Trial; PAC-COPD, Spanish Phenotype and Course of COPD; PLATINO, Proyecto Latinoamericano de Investigación en Obstrucción Pulmonar; SEPOC, Quality of Life of Chronic Obstructive Pulmonary Disease Study.

used to evaluate whether there was a significant difference between AUC values across cohorts for different modeling approaches. Calibration plots were used to show the relationship between model-based predictions of mortality and observed proportions of mortality within prespecified bins. The Brier score (21) was assessed as a combination of discrimination and calibration. The Brier score is the mean squared error for the difference between the predicted probability of death and the true survival status (0 = alive, 1 = dead). Models with smaller Brier score values are preferred. The reference value for the Brier score can be obtained by using the percentage of patients who died in the study population as the individual predicted probability of death.

In addition, we constructed 2 scenarios for which we predicted individual absolute risks of 3-year mortality based on the different approaches. One scenario included a 70-year-old man with COPD, an MRC dyspnea scale grade of 4 (original scale), and an FEV_1 percent predicted of 50% of less who came from a specialty care setting. The other scenario included a 70-year-old woman with COPD, an MRC dyspnea scale grade of 2 (original scale), and an FEV_1 percent predicted greater than 50% who came from nonspecialty care setting. We evaluated the predicted probability of death within 3 years for both scenarios and for the different approaches. We then calculated the percentage changes of these probabilities compared with the predicted probabilities of the benchmark model. All analyses were performed with R for Windows (R Foundation for Statistical Computing, Vienna, Austria) (22) and the R packages lme4 (23), nlme (24), mi (25), MKmisc (26), and PresenceAbsence (27).

RESULTS

The size of the cohorts varied from 106 to 2,273 patients, and the 3-year mortality rate varied from 7% to 34%. The

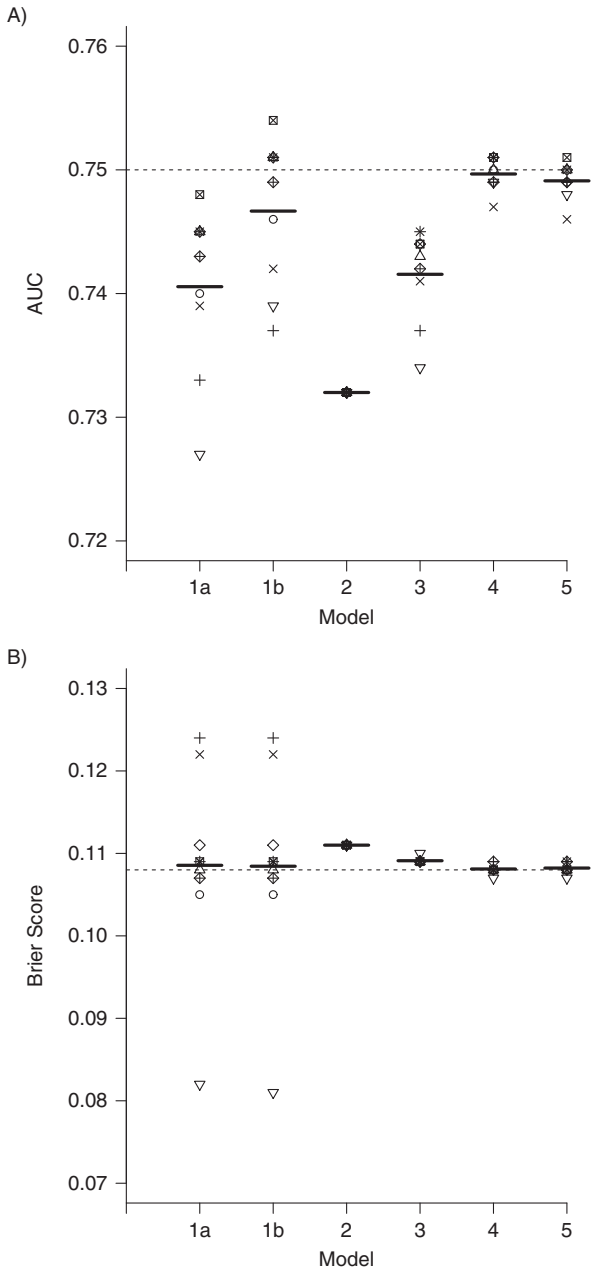

Figure 1. Area under the receiver operating characteristic curve (AUC) (A) and Brier score (B) for each of the 6 modeling options for dealing with a missing predictor and with respect to the individual cohort left out using data from 9 cohort studies. Patients whose data were used for this analysis were recruited between 1981 and 2006. The dashed line indicates AUC or Brier score using the full data set (benchmark). Modeling options 4 and 5 are based on multiple-imputation techniques. The excluded numbers of patients per cohort were: Barmelweid cohort, $n = 231$ (○); Basque study, $n = 106$ (△); Copenhagen City Heart Study, $n = 2,273$ (+); Cardiovascular Health Study, $n = 2,109$ (×); Jackson Heart Study, $n = 419$ (◇); National Emphysema Treatment Trial, $n = 1,934$ (▽); Phenotype and Course of Chronic Obstructive Pulmonary Disease, $n = 330$ (☒);Proyecto Latinoamericano de Investigación en Obstrucción Pulmonar, $n = 173$ (✳); and Quality of Life of Chronic Obstructive Pulmonary Disease Study, $n = 317$ (☼). Modeling options were "exclude study, use study-specific setting" (1a), "exclude study, use random study effects" (1b), "exclude predictor" (2), "mode imputation" (3), "multiple imputation with fixed study effects" (4), and "multiple imputation with random study effects" (5).

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

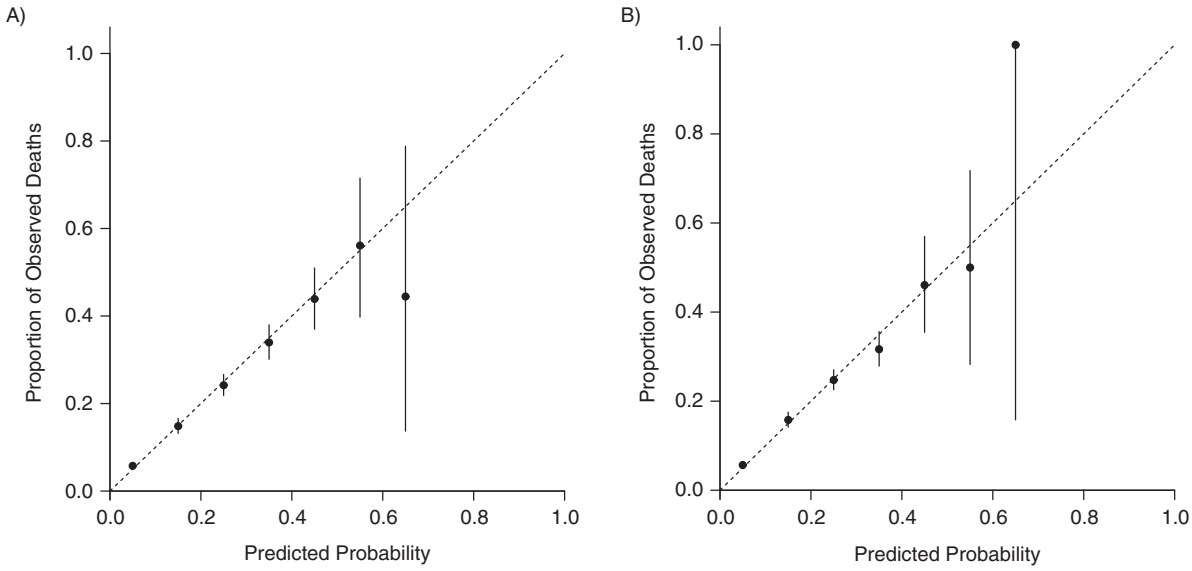

Figure 2. Calibration plots of the benchmark model (A) and the model without the predictor Medical Research Council dyspnea scale grade (B) in 9 cohort studies, 1981–2006. The Figure shows grouping of the predicted versus observed data in 0.1-step bins. The included cohorts were the Barmelweid cohort, Basque study, Copenhagen City Heart Study, Cardiovascular Health Study, Jackson Heart Study, National Emphysema Treatment Trial, Phenotype and Course of Chronic Obstructive Pulmonary Disease, Proyecto Latinoamericano de Investigación en Obstrucción Pulmonar, and Quality of Life of Chronic Obstructive Pulmonary Disease Study. Bars, 95% confidence intervals.

overall 3-year mortality rate was 14%, which resulted in a reference Brier score of 0.120. Table 1 summarizes the cohort-specific 3-year mortality rates. A detailed description of the included cohorts can be found elsewhere (11).

Web Table 1 (available at http://aje.oxfordjournals.org/) shows the summary of model performance as measured using AUC and Brier score, including 95% confidence intervals applied to the full data set and with MRC dyspnea scale grade set to missing for each of the cohorts separately. Our benchmark model—a model with all predictors and random study effects fitted to the full data set—had an AUC of 0.750 (95% confidence interval: 0.735, 0.765) and a Brier score of 0.108 (95% confidence interval: 0.104, 0.113).

Results with respect to AUC and Brier score

A summary of the results with respect to the AUC and Brier score is presented in Figure 1A and 1B. When comparing the AUCs resulting from the different approaches when 1 cohort had a missing MRC dyspnea scale grade for all patients, we obtained the lowest AUC (0.732) for the case in which the predictor MRC dyspnea scale grade was left out of the model. We obtained the best results with respect to average AUC across the 9 cohorts with the fixed-effects model with multiple imputation (AUC between 0.747 and 0.751), although multiple imputation with random effects performed almost as well. Imputing the most frequent value resulted in performance worse than that obtained by excluding the cohort with the missing predictor while including a random study effect. Exclusion of a cohort resulted in much higher variability of the AUC compared with the multiple-imputation options, which may be due to the observed variables rather than deleted variables. We compared the AUC values resulting from modeling approach 1a across the 9 cohorts with those of modeling approach 4 using the Wilcoxon test. The resulting P value was <0.001, demonstrating that multiple imputation with fixed study effects resulted in significantly higher AUC values.

The cohorts from the Copenhagen City Heart Study, National Emphysema Treatment Trial, and Cardiovascular Health Study each included more than 1,900 patients. When 1 of these cohorts was left out, the resulting AUC values were relatively small for modeling options 1a, 1b, and 3. When multiple imputation was used with fixed or random study effects (modeling approaches 4 and 5), the AUC values were much closer to those achieved when small cohorts were left out, indicating that this approach is superior especially when the number of missing values is large.

With respect to the Brier score, we again found that the model in which we left out the entire predictor had the largest Brier score (0.111; 95% confidence interval: 0.106, 0.116), and we obtained lowest Brier scores with fixed or random study effects (values ranging from 0.107 to 0.109 over all cohorts excluded).

Figure 2 presents the calibration of the benchmark model fitted to the full data set and with the model leaving out MRC dyspnea scale grade (modeling approach 2) as a predictor. Figure 2A shows that the benchmark model was very well calibrated. Only in the highest bin did the predicted probabilities overestimate the true proportion of deaths. For the model in which the missing predictor was left out (Figure 2B), calibration was fairly good, but in the highest bin, the predicted probabilities underestimated the true proportion of deaths.

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

Table 2. Predicted 3-Year Mortality in 2 Scenarios Using a Selection of Modeling Approaches, 9 Chronic Obstructive Pulmonary Disease Cohort Studies Conducted in the United States, Europe, and Latin America, 1981–2006

Scenario and Model	Predicted 3-Year Mortality	Deviation (% Points)
Scenario 1[a]		
Benchmark	0.36	
1a[b]	0.40	4
2[c]	0.26	−10
4[d]	0.39	3
5[e]	0.38	2
Scenario 2[f]		
Benchmark	0.12	
1a[b]	0.10	−2
2[c]	0.10	−2
4[d]	0.13	1
5[e]	0.13	1

[a] Male patient who was 70 years of age with a Medical Research Council dyspnea scale grade of 4 and forced expiratory volume in 1 second ≤50% from the Quality of Life of Chronic Obstructive Pulmonary Disease Study cohort.
[b] Cohort left out; reduced data set.
[c] Full data set without the predictor Medical Research Council dyspnea scale grade.
[d] Full data set; multiple imputation with fixed study effects.
[e] Full data set; multiple imputation with random study effects.
[f] Female patient who was 70 years of age with a Medical Research Council dyspnea scale grade of 2 and forced expiratory volume in 1 second >50% from the Copenhagen City Heart Study cohort.

Scenarios

As described above, we chose 2 common scenarios of patient characteristics. The predicted 3-year mortality rate for each of the approaches is displayed in Table 2. For these 2 scenarios, the imputation methods using either fixed study effects (modeling approach 4) or random study effects (modeling approach 5) resulted in predicted 3-year mortality probabilities that were very close to those obtained when the full data pool was included. Again, when the missing predictor was left out, the probability of death was underestimated. These results correspond to those from the calibration plots shown in the Figures.

DISCUSSION

In the present study, our aim was to explore how systematically missing predictors could be imputed and to evaluate how the model's predictive performance changes depending on imputation techniques and cohort-specific characteristics in an IPD meta-analysis. We used data from 9 cohorts comprising participants with COPD that had sample sizes that varied from 106 to 2,273 patients. We set predictor variables to missing cohort-wide, sequentially over each of the 9 cohorts included. We found that eliminating the cohort with the missing predictor (modeling approaches 1a and 1b) led to high variability in discriminative ability and calibration (as expressed in AUC and Brier score). When the variables used with the cohort that was omitted differed substantially from those used with the other cohorts, the AUC was overestimated and the Brier score was underestimated compared with the benchmark AUC and Brier score. When the predictor was eliminated from the model (modeling approach 2), the discriminative ability was the smallest and the Brier score was largest.

We used different approaches to deal with the missing predictor cohort-wide. The simple approach of mode imputation (modeling approach 3) outperformed (on average) the elimination of cohorts with information on specialized/nonspecialized care setting (approach 1a) or predictors (approach 2). More advanced approaches based on multiple imputation of the missing data with fixed (modeling approach 4) or random (modeling approach 5) study effects gave results very close to those from the benchmark model, in which all predictor variables were available in all cohorts. Results of the multiple-imputation approaches also had largest average AUCs, as well as the smallest variability across excluded cohorts. When large cohorts (more than 1,900 patients) were excluded, the simple approaches resulted in low AUC values, whereas the AUCs based on multiple-imputation approaches with fixed or random study effects were much less affected by sample size. Similar results can be seen with respect to the Brier score: The variability is very large for the scenarios in which cohorts are left out (modeling approaches 1a and 1b), whereas variability and bias were lowest when using the multiple-imputation approaches (approaches 4 and 5). We concluded that excluding cohorts leads to higher variability in AUC and Brier score than using imputation methods, and the variability in model performance can be substantial depending on the characteristics of the excluded cohort.

Any prediction model developed in a particular population needs critical evaluation of its performance in other populations. For some prediction models, such as models to predict cardiovascular events, there are many cohort studies in which all of the predictors are available (28). However, many published prediction models are never reevaluated (8), with one of the reasons being the unavailability of 1 or more predictors. For example, Steyerberg (29) described a model for the prediction of abdominal aneurysm mortality that was validated in 2 studies, 1 in the United Kingdom and 1 in the Netherlands. In the UK validation study, 2 predictors were missing, whereas in the Dutch validation study of the same model, only 1 predictor was missing. In the Dutch study, investigators found much better performance than did those in the UK study, indicating that leaving out a predictor may result in underestimation of a model's predictive performance. In a recent paper by Siddique et al. (30), the authors addressed the problem of outcome variables of interest (depression measures) in an IPD meta-analysis being measured in different ways. They considered the situation as a missing-data problem and used a Bayesian approach to multiply impute the missing outcome information while taking into account the relationship between the measures from calibration studies.

In a recent paper by Collins et al. (3), the authors addressed the conduct in and reporting of external validation studies and found that the vast majority of studies were lacking with respect to reporting, design, and handling of missing data. In a

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

case in which a single cohort with a missing predictor is available, we propose incorporating the study setting into the intercept and using multiple imputation. If more than 1 cohort is available, multiple imputation with fixed or random study effects is the best way to comprehensively reevaluate prediction models. Our approach can also be combined with multiple imputation of partially missing data. If a small percentage of values is missing for individual patients, these can be imputed together with the missing predictor. Two-stage multiple imputation (31) may also be used. In the first step, partially missing data are imputed for each cohort; imputation of the entire predictor then follows when necessary.

To our knowledge, the present study is the first in which researchers have addressed the question of model performance as measured with discriminative ability and calibration when leaving out cohorts or using different imputation techniques for cohort-wide missing predictors in a logistic regression setting. We had 9 cohorts with COPD patients that contained all 3 predictors of a published prediction model for 3-year mortality (1), as well as the outcome. We left out the modified MRC dyspnea scale grade across these cohorts and used 6 different methods for handling the cohort-wide missing data. Across the different cohorts, we evaluated model performance for the different modeling approaches, and we compared the findings with the reference categories in which no data were missing (benchmark).

Two of the 6 modeling approaches reflected a poorly performing but common approach in practice: omitting the cohort with the missing predictor. In 1 of these approaches, information on specialized/nonspecialized care setting was included; in the other approach, random study effects were included. The general design of the present study may be used to further explore imputation methods in prediction models that contain more predictors or predictors with other functional forms (e.g., continuous or transformed data). In our study, the multiple-imputation model for the MRC dyspnea scale grade predictor included the variables of death, age, FEV_1, sex, and study. This set of variables is not specifically and closely related to the missing predictor, but it does contain a variable outside of the prediction model (sex). The potential of multiple-imputation approaches for making a validation cohort suitable for the predictive model of interest could be even larger if different but similar variables need to be imputed. This could, for example, include a biomarker measured in blood samples as opposed to urine samples.

Our study has several strengths and limitations. Strengths of our study include the number and diversity of cohorts in terms of sample size, setting, disease severity, and outcome frequency (disease mix), which allowed us to evaluate the consistency of a particular imputation technique. Also, we used different measures for model performance, including AUC, Brier score, and calibration plots. For multiple imputation, we chose $m = 5$ imputations, but $m = 3$ could also be sufficient if less than 20% of data are missing (20, 32). In more recent papers, it has been suggested that the number of imputations should be higher with higher percentages of missing values (33). Across the cohorts we analyzed, the percentage of missing values ranged from 1.3% to 28.8%.

Our study has several limitations. We focused on a single prediction model with just 3 predictors and 3-year mortality rate as the outcome in patients with COPD. This scenario represents a relatively simple case and calls for extension to other disease areas, types and numbers of missing predictors, and outcomes. Because the outcome variable in our study was binary, we focused on logistic regression approaches. Other approaches include classification trees or neural networks (34), which also have high discriminative power. However, the advantage of logistic regression models is that estimated coefficients can be interpreted easily from a clinical perspective (35). In the present study, we focused on missing data for a single predictor that was missing cohort-wide. It is straightforward to generalize the approach to missing predictors in more than 1 cohort or to missing observations for other variables in the prediction model.

In summary, we explored different methods to impute an uncollected variable for validation of a prediction model. We found that multiple imputation with fixed or random study effects is the best approach to impute values for a predictor for the whole cohort. These approaches consistently outperformed other methods across data sets and metrics for model performance. Results of this study may facilitate the use of cohort studies that do not include all predictors and pave the way for more widespread validation of prediction models.

ACKNOWLEDGMENTS

Author affiliations: Horten Centre for Patient-Oriented Research and Knowledge Transfer, University of Zurich, Zurich, Switzerland (Ulrike Held); Clinical Epidemiology and Medical Technology Assessment, University Hospital Maastricht, Maastricht, the Netherlands (Alfons Kessels); ISGlobal, Center for Research and Environmental Epidemiology, Barcelona, Spain (Judith Garcia Aymerich, Xavier Basagaña); IMIM (Hospital del Mar Research Institute), Barcelona, Spain (Judith Garcia Aymerich, Xavier Basagaña); Universitat Pompeu Fabra, Barcelona, Spain (Judith Garcia Aymerich, Xavier Basagaña); Consorcio de Investigación Biomédica en Red de Epidemiología y Salud Pública, Barcelona, Spain (Judith Garcia Aymerich, Xavier Basagaña); Department of General Practice, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (Gerben ter Riet); Julius Centre for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands (Karel G. M. Moons); Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland (Milo A. Puhan); and Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Milo A. Puhan).

Conflict of interest: none declared.

REFERENCES

1. Puhan MA, Garcia-Aymerich J, Frey M, et al. Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated BODE index and the ADO index. *Lancet*. 2009;374(9691):704–711.
2. Steyerberg EW. Dealing with missing values. In: Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer; 2009:115–137.

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016

3. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40.
4. Siddique J, Harel O, Crespi CM. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *Ann Appl Stat*. 2012;6(4):1814–1837.
5. Burgess S, White IR, Resche-Rigon M, et al. Combining multiple imputation and meta-analysis with individual participant data. *Stat Med*. 2013;32(26):4499–4514.
6. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012; 367(14):1355–1360.
7. Li T, Hutfless S, Scharfstein DO, et al. Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. *J Clin Epidemiol*. 2014;67(1):15–32.
8. Ahmed I, Debray TP, Moons KG, et al. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol*. 2014;14:3.
9. Resche-Rigon M, White IR, Bartlett JW, et al. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med*. 2013; 32(28):4890–4905.
10. Jolani S, Debray TP, Koffijberg H, et al. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*. 2015;34(11):1841–1863.
11. Puhan MA, Hansel NN, Sobradillo P, et al. Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. *BMJ Open*. 2012;2(6):e002152.
12. Sobradillo P, Iriberri M, Gomez B, et al. Validation of bode index as a predictor of mortality in COPD patients. *Eur Respir J*. 2008;32(suppl 52):531.
13. Appleyard M, Hansen A, Schnohr P. The Copenhagen City Heart Study: a book of tables with data from the first examination (1976–78) and a five years follow-up (1981–1983). *Scand J Soc Med*. 1989;170(suppl 41):1–160.
14. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol*. 1991;1(3): 263–276.
15. Carpenter MA, Crow R, Steffes M, et al. Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. *Am J Med Sci*. 2004;328(3): 131–144.
16. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. 2003;348(21): 2059–2073.
17. Garcia-Aymerich J, Gomez FP, Anto JM. Phenotypic characterization and course of chronic obstructive pulmonary disease in the PAC-COPD Study: design and methods. *Arch Bronconeumol*. 2009;45(1):4–11.
18. Menezes AM, Perez-Padilla R, Jardim JR, et al. Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *Lancet*. 2005;366(9500): 1875–1881.
19. Domingo-Salvany A, Lamarca R, Ferrer M, et al. Health-related quality of life and mortality in male patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2002;166(5):680–685.
20. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681–694.
21. Bradley AA, Schwartz SS, Hashino T. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather Forecast*. 2008;23(5):992–1006.
22. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
23. Bates D, Maechler M, Bolker B, et al. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1–48.
24. Pinheiro J, Bates D, DebRoy S, et al. nlme: linear and nonlinear mixed effects models. 2015. http://CRAN.R-project.org/package=nlme. Updated May 10, 2016. Accessed December 1, 2015.
25. Gelman A, Hill J, Su YS, et al. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Softw*. 2011;45(2):1–31.
26. Kohl M. MKmisc: miscellaneous functions from M. Kohl. 2015. https://cran.r-project.org/web/packages/MKmisc/index.html. Updated January 6, 2016. Accessed December 1, 2015.
27. Freeman E, Moisen G. PresenceAbsence: an R package for presence absence analysis. *J Stat Softw*. 2008;23(11):1–31.
28. Eichler K, Puhan MA, Bachmann LM. The role of statins in primary prevention of cardiovascular disease. *Arch Intern Med*. 2007;167(10):1100.
29. Steyerberg EW. Validation of prediction models. In: Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer; 2009:299–311.
30. Siddique J, Reiter JP, Brincks A, et al. Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Stat Med*. 2015; 34(26):3399–3414.
31. Harel O. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Stat Methodol*. 2007;4:75–89.
32. Vergouwe Y, Royston P, Moons KG, et al. Development and validation of a prediction model with missing predictor data: a practical approach. *J Clin Epidemiol*. 2010;63(2):205–214.
33. Reiter JP. Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Stat Probab Lett*. 2008;78(1):15–20.
34. Waljee AK, Higgins PD, Singal AG. A primer on predictive models. *Clin Transl Gastroenterol*. 2014;5:e44.
35. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform*. 2002;35(5-6):352–359.

Downloaded from http://aje.oxfordjournals.org/ at Universitaet Zuerich on September 21, 2016